The Role of Time in Music Emotion Recognition: Modeling Musical Emotions from Time-Varying Music Features

Marcelo Caetano[1]*, Athanasios Mouchtaris[1,2], and Frans Wiering[3]

[1] Institute of Computer Science, Foundation for Research and Technology - Hellas FORTH-ICS, Heraklion, Crete, Greece
[2] University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-71409
[3] Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands
caetano@ics.forth.gr, mouchtar@ics.forth.gr, f.wiering@uu.nl

Abstract. Music is widely perceived as expressive of emotion. However, there is no consensus on which factors in music contribute to the expression of emotions, making it difficult to find robust objective predictors for music emotion recognition (MER). Currently, MER systems use supervised learning to map non time-varying feature vectors into regions of an emotion space guided by human annotations. In this work, we argue that time is neglected in MER even though musical experience is intrinsically temporal. We advance that the temporal variation of music features rather than feature values should be used as predictors in MER because the temporal evolution of musical sounds lies at the core of the cognitive processes that regulate the emotional response to music. We criticize the traditional machine learning approach to MER, then we review recent proposals to exploit the temporal variation of music features to predict time-varying ratings of emotions over the course of the music. Finally, we discuss the representation of musical time as the flow of musical information rather than clock time. Musical time is experienced through auditory memory, so music emotion recognition should exploit cognitive properties of music listening such as repetitions and expectations.

Keywords: Music, Time, Emotions, Mood, Automatic Mood Classification, Music Emotion Recognition

1 Introduction

One of the recurring themes in treatises of music is that music both evokes emotions in listeners (emotion induction) and expresses emotions that listeners perceive, recognize, or are moved by, without necessarily feeling the emotion

* This work is funded by the Marie Curie IAPP "AVID MODE" grant within the European Commissions FP7.

(emotion perception) [14]. The emotional impact of music on people and the association of music with particular emotions or 'moods' have been used in certain contexts to convey meaning, such as in movies, musicals, advertising, games, music recommendation systems, and even music therapy, music education, and music composition, among others. Empirical research on emotional expression started about one hundred years ago, mainly from a music psychology perspective [9], and has successively increased in scope up to today's computational models. Research on music and emotions usually investigates listeners' response to music by associating certain emotions to particular pieces, genres, styles, performances, among many others.

The mechanisms whereby music elicits emotions in listeners are not well understood. A central question in the study of music and emotions is "Which attributes or musical qualities, if any, elicit emotional reactions in listeners? [14, 31]" At first, we should identify factors in the listener, in the music, and in the context that influence musical emotions (i.e., emotional reactions to music). Only then can we proceed to develop a theory about specific mechanisms that mediate among musical events and experienced emotions.

Among the causal factors that potentially affect listeners' emotional response to music are *personal, situational,* and *musical*. Personal factors include age, gender, personality, musical training, music preference, and current mood. Situational factors can be physical such as acoustic and visual conditions, time and place, or social such as type of audience, and occasion. Musical factors include genre, style, key, tuning, orchestration, among many others.

Juslin and Västfjäll [14] sustain that there is evidence of emotional reactions to music in terms of various subcomponents, such as *subjective feeling, psychophysiology, brain activation, emotional expression, action tendency, emotion regulation* and these, in turn, feature different psychological mechanisms like *brain stem reflexes, evaluative conditioning, emotional contagion, visual imagery, episodic memory, rhythmic entrainment,* and *musical expectancy*. They state that "none of the mechanisms evolved for the sake of music, but they may all be recruited in interesting (and unique) ways by musical events. Each mechanism is responsive to its own combination of information in the music, the listener, and the situation."

The literature on the emotional effects of music [15, 9] has accumulated evidence that listeners often agree about the emotions expressed (or elicited) by a particular piece, suggesting that there are aspects in music that can be associated with similar emotional responses across cultures, personal bias or preferences. Several researchers imply that there is a causal relationship between music features and emotional response [9], giving evidence that certain music dimensions and qualities communicate similar affective experiences to many listeners.

An emerging field is the automatic recognition of emotions (or 'mood') in music, also called music emotion recognition (MER) [17]. The aim of MER is to design systems to automatically estimate listeners' emotional reactions to music. A typical approach to MER categorizes emotions into a number of classes and applies machine learning techniques to train a classifier and compare the re-

sults against human annotations [17, 49, 23]. The 'automatic mood classification' task in MIREX epitomizes the machine learning approach to MER, presenting systems whose performance range from 22 to 65 percent [11]. Some researchers speculate that musical sounds can effectively cause emotional reactions (via *brain stem reflex*, for example). Researchers are currently investigating [12, 17] how to improve the performance of MER systems. Interestingly, the role of time in the automatic recognition of emotions in music is seldom discussed in MER research.

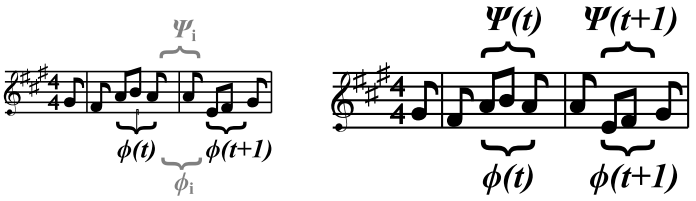
Musical experience is inherently tied to time. Studies [19, 24, 13, 36] suggest that the temporal evolution of the musical features is intrinsically linked to listeners' emotional response to music, that is, emotions expressed or aroused by music. Among the cognitive processes involved in listening to music, memory and expectations play a major role. In this article, we argue that time lies at the core of the complex link between music and emotions, and should be brought to the foreground of MER systems.

The next section presents a brief review of the classic machine learning approach to MER. We present the traditional representation of musical features and the model of emotions to motivate the incorporation of temporal information in the next section. Then, we discuss an important drawback of this approach, the lack of temporal information. The main contribution of this work is the detailed presentation of models that exploit temporal representations of music and emotions. We also discuss modeling the relationship between the temporal evolution of musical features and emotional changes. Finally, we speculate on different representations of time that better capture the experience of musical time before presenting the conclusions and discussing future perspectives.

2 Machine Learning and Music Emotion Recognition

Traditionally, computational systems that automatically estimate the listener's emotional response to music use *supervised learning* to train the system to map a feature space representing the music onto a model of emotion according to annotated examples [17, 49, 23, 11]. The system can perform classification [21] or regression [48], depending on the nature of the representation of emotions (see section 2.2). After training, the system can be used to predict listeners' emotional responses to music that was not present in the training phase, assuming that it belongs to the same data set and therefore can be classified under the same underlying rules. System performance is measured comparing the output of the system with the annotation for the track.

Independently of the specific algorithm used, the investigator that chooses this approach must decide how to represent the two spaces, the music features and the emotions. On the one hand, we should choose music features that capture information about the expression of emotions. Some features such as tempo and loudness have been shown to bear a close relationship with the perception of emotions in music [38]. On the other hand, the model of emotion should reflect listeners' emotional response because emotions are very subjective and may change according to musical genre, cultural background, musical training



(a) Bag-of-Features (b) Time Series

Fig. 1. Illustration of feature extraction. Part a) shows the bag-of-features approach, where the music piece is represented by a non time-varying vector of features Φ_i averaged from successive frames. Notice that there is only one global emotion Ψ_i associated with the entire piece as well. In part b), Both music features Φ and emotion annotations Ψ are kept as a time series.

and exposure, mood, physiological state, personal disposition and taste [9]. We argue that the current approach misrepresents both music and listeners' emotional experience by neglecting the role of time. In this article, we advance that the temporal variation of music features rather than the feature values should be used as predictors of musical emotions.

2.1 Music Features

Typically, MER systems represent music with a vector of features. The features can be extracted from different representations of music, such as the audio, lyrics, the score, social tags, among others [17]. Most machine learning methods described in the literature use the audio to extract the music features [17, 49, 23, 11]. Music features such as root mean square (RMS) energy, mel frequency cepstral coefficients (MFCCs), attack time, spectral centroid, spectral rolloff, fundamental frequency, and chromagram, among many others, are calculated from the audio by means of signal processing algorithms [27, 12, 48]. The number and type of features dictates the dimensionality of the input space (some features such as MFCCs are multidimensional). Therefore, there usually is a feature selection or dimensionality reduction step to determine a set of uncorrelated features. A common choice for dimensionality reduction is principal component analysis (PCA)[26, 12, 21]. Huq *et. al* [12] investigate four different feature selection algorithms and their effect on the performance of a traditional MER system. Kim *et. al* [17] presented a thorough state-of-the-art review of MER in 2010, exploring a wide range of research in MER systems, particularly focusing on methods that use textual information (e.g., websites, tags, and lyrics) and content-based approaches, as well as systems combining multiple feature domains (e.g., features plus text). Their review is evidence that MER systems rarely exploit temporal information.

The term 'semantic gap' has been coined to refer to perceived musical information that does not seem to be contained in the acoustic patterns present in the audio, even though listeners agree about its existence [47]. Music happens

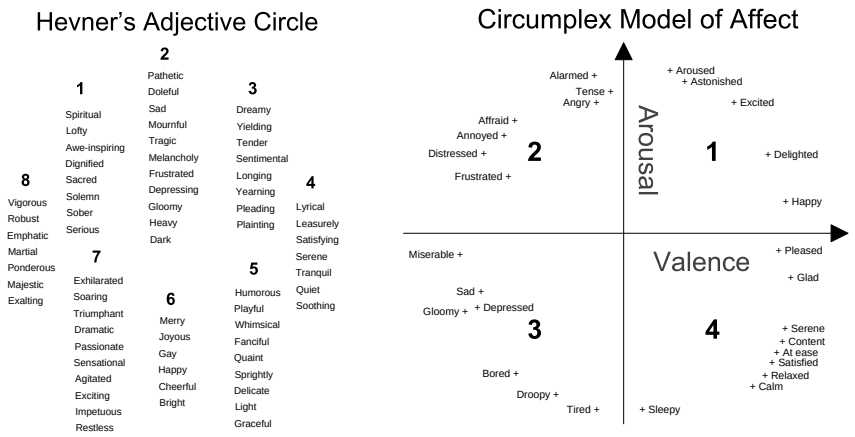
essentially in the brain, so we need to take the cognitive mechanisms involved in processing musical information into account if we want to be able to model people's emotional response to music. Low-level audio features give rise to high-level musical features in the brain, and these, in turn, influence emotion recognition (and experience). This is where we argue that time has a major role, still neglected in most approaches found in the literature. However, only very recently have researchers started to investigate the role of time in MER. On the one hand, the different time scales in musical experience should be respected [42]. On the other hand, the temporal changes of some features are more relevant than feature values isolated from the musical context [3].

Usually, MER systems use a "bag of features" approach, where all the features are stacked together [12]. However, these features are associated with different levels of music experience, namely, the *perceptual*, the *rhythmic*, and the *formal* levels. These levels, in turn, are associated with different time scales [42]. Music features such as *pitch*, *loudness*, and *duration* are extracted early in the processing chain that converts sound waves reaching the ear into sound perception in the brain. Rhythm and melody depend hierarchically on the features from the previous level. For example, melody depends on temporal variations of pitch. Subsequently, the formal level is comprised of structural blocks from the melodic and harmonic level.

Figure 1 illustrates the music feature extraction step in MER. Typically, these features are calculated from successive frames taken from excerpts of the audio that last a few seconds [17, 49, 23, 11, 12] and then averaged like seen in part a) of figure 1, losing the temporal correlation [23]. Consequently, the whole piece (or track) is represented by a static (non time-varying) vector, intrinsically assuming that musical experience is static and that the listener's emotional response can be estimated from the audio alone. Notice that, typically, each music piece (or excerpt) is associated with only one emotion, represented by Ψ_i in figure 1. The next section explores the representation of emotions in more detail.

2.2 Representation of Emotions

The classification paradigm of MER research uses categorical descriptions of emotions where the investigator selects a set of "emotional labels" (usually mutually exclusive). Part a) of figure 2 illustrates these emotional labels (Hevner's adjective circle [10]) clustered in eight classes. The annotation task typically consists of asking listeners to choose a label from one of the classes for each track. The choice of the emotional labels is important and might even affect the results. For example, the terms associated with music usually depend on genre (pop music is much more likely than classical music to be described as "cool"). As Yang [49] points out, the categorical representation of emotions faces a granularity issue because the number of classes might be too small to span the rich range of emotions perceived by humans. Increasing the number of classes does not necessarily solve the problem because the language used to categorize emotions is ambiguous and subjective [9]. Therefore, some authors [17, 49] have proposed to



(a) Categorical (b) Parametric

Fig. 2. Examples of models of emotion. The left-hand side shows Hevner's adjective circle [10], a categorical description. On the right, we see the circumplex model of affect [30], a parametric model.

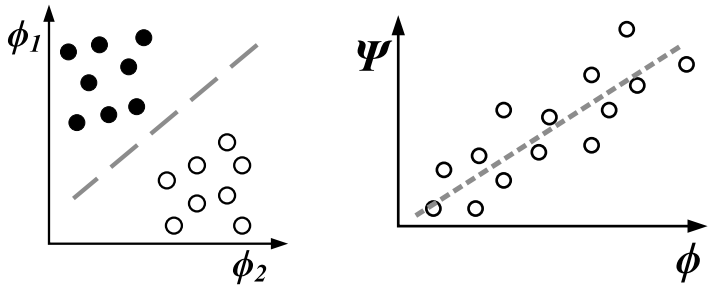
adopt a parametric model from psychology research [30] known as the circumplex model of affect (CMA). The CMA consists of two independent dimensions whose axes represent continuous values of valence (positive or negative semantic meaning) and arousal (activity or excitation). Part b) of figure 2 shows the CMA and the position of some adjectives used to describe emotions associated with music in the plane. An interesting aspect of parametric representations such as the CMA lies in the continuous nature of the model and the possibility to pinpoint where specific emotions are located. Systems based on this approach train a model to compute the valence and arousal values and represent each music piece as a point in the two-dimensional emotion space [49].

One common criticism of the CMA is that the representation does not seem to be metric. That is, emotions that are very different in terms of semantic meaning (and psychological and cognitive mechanisms involved) can be close in the plane. In this article, we argue that the lack of temporal information is a much bigger problem because music happens over time and the way listeners associate emotions with music is intrinsically linked to the temporal evolution of the musical features. Also, emotions are dynamic and have distinctive temporal profiles (boredom is very different from astonishment in this respect, for example).

2.3 Mathematical Notation

In mathematical terms, the traditional approach to MER models the relationship between music $\mathbf{\Phi}$ and emotions $\mathbf{\Psi}$ according to the following

$$\mathbf{\Psi} = f\left(\mathbf{\Phi}, A, \epsilon\right) \tag{2.1}$$



(a) Classification (b) Regression

Fig. 3. Simple examples of machine learning applied to music emotion recognition. Part a) shows an example of classification. In part b), we see an example of regression.

where $\mathbf{\Psi}$ represents the emotion space, $\mathbf{\Phi}$ represents the music, f models the functional relationship between $\mathbf{\Phi}$ and $\mathbf{\Psi}$ parameterized by A with error ϵ. Therefore, in this approach, MER becomes finding the values for the parameters $A = \{a_0, a_1, ..., a_N\}$ that minimize the error $\{\epsilon\}$ and correctly map each $\Phi_i \in \mathbf{\Phi}$ onto their corresponding $\Psi_i \in \mathbf{\Psi}$. Notice that subscript i means an instance of the pair $\{\Psi, \Phi\}$ (an annotated music track). Here, $\Phi_i = [\phi_1, \phi_2, ..., \phi_N]$ is an N dimensional vector of music features and Ψ_i can be a semantic label representing an emotion for the classification case or continuous values of psychological models such as a valence/arousal pair $\Psi_i = \{\upsilon, \alpha\}$.

Figure 3 shows a simple example of classification and regression to illustrate equation (2.1). Part a) illustrates linear classification into two classes, while part b) shows linear regression. In part a), the black dots represent instances of the first class, while the white dots represent the other class. The dashed line is the linear classifier (i.e., the MER system) that separates the input parameter space $\mathbf{\Phi} = \{\phi_1, \phi_2\}$ into two regions that correspond to the classes $\mathbf{\Psi} = \{black, white\}$. For example, a MER system that takes chords as input and outputs the label *happy* for major chords and *sad* for minor chords. In this case, $\mathbf{\Phi}$ is *major* or *minor* and could be encoded as ϕ_1 the first interval and ϕ_2 the second interval in cents, f is a binary classifier (such as a straight line with parameters $A = \{a_0, a_1\}$), and $\mathbf{\Psi} = \{happy, sad\}$. The error ϵ would be associated with misclassification, that is, points associated with one class by the system but labeled with the other. The system could be then used to classify inputs (music) that were not a part of the training data into "happy" or "sad" depending on which category (region) it falls into.

Part b) shows Ψ as a linear function of a single variable ϕ as $\Psi = a_0 + a_1\phi$. In this case, the dots are values of the independent variable or predictor ϕ associated with Ψ. For instance, ϕ represents loudness values positively correlated with arousal, represented by Ψ. Notice that both ϕ and Ψ are real-valued, and the MER system f modeling the relationship between them is the straight dashed line with parameters $A = \{a_0, a_1\}$ obtained by regression (expectation maxi-

mization or least-squares). The modeling error ϵ being minimized is the difference between the measures (the dots in the figure) and the model (the dashed line). The MER system can estimate *arousal* for new music tracks solely based on *loudness* values.

A more general MER system following the same approach would model Ψ as a linear combination of predictors Φ using multiple regression as follows

$$\Psi_i = a_0 + a_1\phi_{i,1} + ... + a_N\phi_{i,N} + ... + \epsilon \tag{2.2}$$

where Ψ_i is the representation of emotion and $\Phi_i = \{\phi_{i,n}\}$ are the music features. This model assumes that emotions can be estimated as a linear combination of the music features, such as $\Phi_i = \{\text{loud}, \text{fast}\}$ music is considered $\Psi = \{\text{upbeat}\}$. Generally, the errors ϵ are supposed uncorrelated with one another (additive error) and with $\mathbf{\Phi}$, whose underlying *probability distribution* has a major influence on the parameters A. Naturally, fitting a straight line to the data is not the only option. Sophisticated machine learning algorithms are usually applied to MER, such as support vector machines [12, 17]. However, these algorithms are seldom appropriate to deal with the temporal nature of music and the subjective nature of musical emotions.

2.4 Where Does the Traditional Approach Fail?

The traditional machine learning approach to MER assumes that the music features are good predictors of musical emotions due to a causal relationship between $\mathbf{\Phi}$ and $\mathbf{\Psi}$. The map from feature space to emotion space is assumed to implicitly capture the underlying psychological mechanisms leading to an emotional response in the form of a one-to-one relationship. However, psychological mechanisms of emotional reactions to music are usually regarded as information processing devices at various levels of the brain, using distinctive types of information to guide future behavior. Therefore, even when the map f explains most of the correlation between between $\mathbf{\Phi}$ and $\mathbf{\Psi}$, it does not necessarily mean that it captures the underlying psychological mechanism responsible for the emotional reaction (i.e., correlation does not imply causation). In other words, while equation (2.1) can be used to model the relationship between music features and emotional response, it does not imply the existence of causal relations between them.

Equation (2.1) models the relationship between music features and emotional response from a behavioral viewpoint, supposing that the emotional response is consistent across listeners, irrespective of cultural and personal context. Currently, MER systems rely on self-reported annotations of emotions using a model such as Hevner's adjective circle or the CMA. On the one hand, this approach supposes that the model of emotion allows the expression of a broad palette of musical emotions. On the other hand, it supposes that self-reports are enough to describe the outcome of several different psychological mechanisms responsible for musical emotions [14]. Finally, the listener's input is only provided in the form of annotations and only used when comparing these annotations to

the emotional labels output by the system, neglecting *personal* and *situational* factors. The terms semantic gap [47, 4] and 'glass ceiling' [1] have been coined to refer to perceived musical information that does not seem to be contained in the audio even though listeners agree about its existence. MER research needs to bridge the gap between the purely acoustic patterns of musical sounds and the emotional impact they have on listeners by modeling the generation of musical meaning [15]. Musical experience is greater than auditory impression [22]. The so called 'semantic gap' is a mere reflection of how the current typical approach misrepresents both the listener and musical experience.

Here we argue that the current approach misrepresents both music and listeners' emotional experience by neglecting the role of time. Currently, MER research ignores evidence [19, 24, 13, 14] suggesting the existence of complex relationships between the dynamics of musical emotions and the response to how musical structure unfolds in time. The examples given in figure 3 illustrate this point (although in a very simplified way). Neither system uses temporal information at all. In part a), the input music is classified as "happy" or "sad" based solely on whether it uses major or minor chords, ignoring chord progression, inversions, etc. Part b) supposes a rigid association between *loudness* and *arousal* (loud music is arousing), ignoring temporal variations (like sudden changes from soft to loud).

Krumhansl [20] suggests that music is an important part of the link between emotions and cognition. More specifically, Krumhansl investigated how the dynamic aspect of musical emotion relates to the cognition of musical structure. According to Krumhansl, musical emotions change over time in intensity and quality, and these emotional changes covary with changes in psycho-physiological measures [20]. Musical meaning and emotion depend on how the actual events in the music play against this background of expectations. David Huron [13] wrote that humans use a general principle in the cognitive system that regulates our expectations to make predictions. According to Huron, music (among other stimuli) influences this principle, modulating our emotions. Time is a very important aspect of musical cognitive processes. Music is intrinsically temporal and we need to take into account the role of human memory when experiencing music. In other words, musical experience is learned. As the music unfolds, the learned model is used to generate expectations, which are implicated in the experience of listening to music. Meyer [25, 24] proposed that expectations play the central psychological role in musical emotions.

3 Time and Music Emotion Recognition

We can incorporate temporal information into the representation of the music features and into the emotional response. In the first case we calculate the music features sequentially as a time-series, while the last case consists of recording listeners' annotations of emotional responses over time and keeping the information as a time-series. Figure 1(b) illustrates the music features and emotions associated with music (represented by the score) over time. Thus $\phi(t)$ is the

current value of a music feature, and $\phi\left(t+1\right)$ is the subsequent value. Similarly, $\Psi\left(t\right)$ and $\Psi\left(t+1\right)$ follow each other.

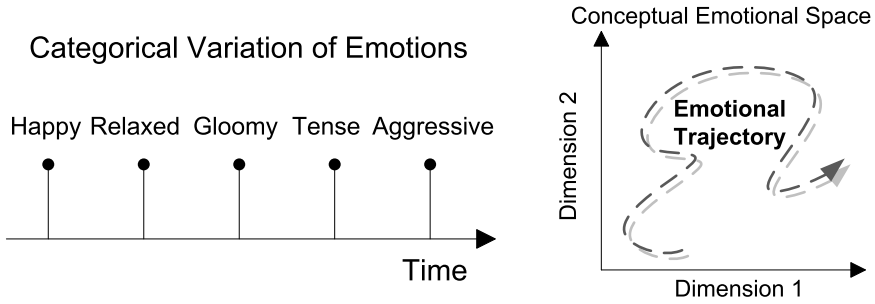
There are several ways of exploiting the information from the temporal variation of music features and emotions. A very straightforward way would be to use time-series analysis and prediction techniques, such as using previous values to predict future values of the series. In this case, the investigator could use past values of a series of valence/arousal $\{\upsilon, \alpha\}$ annotations over time to predict the next $\{\upsilon, \alpha\}$ value. A somewhat more complex approach is to use the temporal behavior of one time series as predictors of the next value of another series. In this case, the temporal variation of the music features would be used as predictors in regression. Thus variations in *loudness* rather than *loudness* values are used to predict the *arousal* associated. Several techniques can be employed, such as regression analysis, dynamical system theory, as well as machine learning algorithms developed to model the dynamic behavior of time series. Thus next section reviews approaches to MER that use the temporal variation of music features as predictors of musical emotions.

3.1 Time Series and Prediction

The feature vector should be calculated for every frame of the audio signal and kept as a time series as shown in figure 1(b). In other words, the music features Φ_i are now represented by a time-varying vector $\Phi_i\left(t\right) = \{\phi_i\left(t\right), \phi_i\left(t-1\right), \phi_i\left(t-2\right), ..., \phi_i\left(t-N\right)\}$. The temporal correlation of the features must be exploited and fed into the model of emotions to estimate listeners' response to the repetitions and the degree of "surprise" that certain elements might have [38]. The simplest way to incorporate temporal information from the music features is to include time differences, such as *loudness* values and also *loudness* variations (from the previous value). This MER system uses information about how loud a certain passage sounds and also if the music is getting louder (building up tension, for example), using previous values of features to predict the next (is *loudness* going to increase or decrease?) and compare these predictions against how the same features are unfolding in the music as follows

$$\phi_i\left(t+1\right) = a_1\phi_i\left(t\right) + a_2\phi_i\left(t-1\right) + a_3\phi_i\left(t-2\right) + ... + \epsilon \tag{3.1}$$

where $\phi_i\left(t+1\right)$ represents the next value for the feature ϕ_i, $\phi_{i,t}$ the present value, $\phi_{i,t-1}$ the previous, and so forth. The predictions $\phi_i\left(t+1\right)$ can be used to estimate listeners' emotional responses. Listeners have expectations about how the music is unfolding in time. For instance, expectations about the next term in a sequence (the next chord in chord progression or the next pitch in melodic contour) or expectations about continuous parameters (become louder or brighter). Whenever listeners' expectations are correct it is rewarding (fulfillment) and when they are not it is unrewarding (tension).



(a) Categorical (b) Continuous

Fig. 4. Temporal variation of emotions. The left-hand side shows emotional labels recorded over time. On the right, we see a continuous conceptual emotional space with an emotional trajectory (time is represented by the arrow).

3.2 Emotional Trajectories

A very simple way of recording information about the temporal variation of emotional perception of music would be to ask listeners to write down the emotional label and a time stamp as the music unfolds. The result is illustrated in figure 4(a). However, this approach suffers from the granularity and ambiguity issues inherent of using a categorical description of emotions. Ideally, we would like to have an estimate of how much a certain emotion is present at a particular time. Krumhansl [19] proposes to collect listener's responses continuously while the music is played, recognizing that retrospective judgments are not sensitive to unfolding processes. However, in this study [19], listeners assessed only one emotional dimension at a time. Each listener was instructed to adjust the position of a computer indicator to reflect how the amount of a specific emotion (for example, sadness) they perceived changed over time while listening to excerpts of pieces chosen to represent the emotions [19].

Recently, there have been proposals to collect self-report of emotional reactions to music [39], including software such as EmotionSpace Lab [35], EmuJoy [28], and MoodSwings [16]. EmotionSpace Lab [35] allows listeners to continuously rate emotions while listening to music as points on the $\{v, \alpha\}$ (valence-arousal) plane (CMA), giving rise to an *emotional trajectory* on a two-dimensional model of emotion like the one shown in figure 4(b) (time is represented by the arrow). Use of the CMA accommodates a wide range of emotional states in a compact representation. Similarly, EmuJoy[28] allows continuous self-report of emotions over time in two-dimensional space (CMA). MoodSwings [16] is an online collaborative game designed to collect second-by-second labels for music using the CMA as model of emotion. The game was designed to capture $\{v, \alpha\}$ pairs dynamically (over time) to reflect emotion changes in synchrony with music and also to collect a distribution of labels across multiple players for a given song or even a moment within a song. Kim *et al.* state that the method provides

quantitative labels that are well-suited to computational methods for parameter estimation.

A straightforward way of using information from the sequence of emotional labels $\Psi_i(t)$ to predict future values would be to use the underlying dynamics of the temporal variation of the sequence itself, like expressed below

$$\Psi_i(t+1) = a_0 + a_1\Psi_i(t) + a_2\Psi_i(t-1) + a_3\Psi_i(t-2) + ... + \epsilon. \tag{3.2}$$

Notice that equation (3.2) fits a linear prediction model to the time series of emotional labels $\Psi_i(t)$ under the assumption that the previous values in the series can be used to predict future values, indicating trends and modeling the inertia of the system. In other words, the model assumes that increasing values of $\Psi_i(t)$ indicate that the next value will continue to increase by a rate estimated from previous rates of growth, for example.

3.3 Modeling Musical Emotions from Time-Varying Music Features

Finally, we should investigate the relationship between the temporal variation of musical features and the emotional trajectories. MER systems should include information about the rate of temporal change of musical features. For example, we should investigate how changes in loudness correlate with the expression of emotions. Early studies used time series analysis techniques to investigate musical structure. Vos *et. al* [46] tested the structural and perceptual validity of notated meter applying autocorrelation to to the flow of melodic internals between notes from thirty fragments of compositions for solo instruments by J. S. Bach.

Recently, researchers started exploring the temporal evolution of music by treating the sequence of music features as a time series modeled by ordinary least squares [36, 38], linear dynamical systems such as Kalman filters [32–34], dynamic texture mixtures (DTM) [8, 44], auto-regressive models (linear prediction) [18], neural networks [5–7, 45], among others. Notice that these techniques are intimately related. For example, the Kalman filter is based on linear dynamical systems discretized in the time domain and modeled as a Markov chain, whereas the hidden Markov model can be viewed as a specific instance of the state space model in which the latent variables are discrete.

First of all, it is important to distinguish between stationary and nonstationary sequential distributions. In the stationary case, the data evolves in time, but the distribution from which it is generated remains the same. For the more complex nonstationary situation, the generative distribution itself is evolving in time.

Ordinary Least Squares Schubert [36, 38] studied the relationship between music features and perceived emotion using continuous response methodology and time-series analysis. In these studies, both the music features $\Phi_n(t)$ and the emotional responses $\Psi_m(t)$ are multidimensional time series. For example,

$\Phi_1(t) = \left[\phi_1(t)\ \phi_1(t-1) \ldots \phi_1(t-N)\right]^T$ are *loudness* values over time and $\Psi_\alpha(t) = \left[\alpha(t)\ \alpha(t-1) \ldots \alpha(t-N)\right]^T$ are arousal ratings annotated over time. Schubert [36, 38] proposes to model each component of $\Psi(t)$ as a linear combination of features $\Phi(t)$ plus a residual error $\epsilon(t)$ as follows

$$\begin{bmatrix} \upsilon(t) \\ \upsilon(t-1) \\ \vdots \\ \upsilon(t-M) \end{bmatrix} = \begin{bmatrix} \phi_1(t) & \phi_2(t) & \ldots & \phi_N(t) \\ \phi_1(t-1) & \phi_2(t-1) & \ldots & \phi_N(t-1) \\ \vdots & \vdots & & \\ \phi_1(t-M) & \phi_2(t-M) & \ldots & \phi_N(t-M) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} + \begin{bmatrix} \epsilon(t) \\ \epsilon(t-1) \\ \vdots \\ \epsilon(t-N) \end{bmatrix} \tag{3.3}$$

where the model parameters $A = \{a_j\}$ are fit so as to best explain variability in $\Psi(t)$. The error term $\epsilon(t)$ is included to account for discrepancies between the deterministic component of the equation and the actual data value. Two fundamental premises of this model are that the error term be reasonably small and that it fluctuate randomly. Notice that the error term $\epsilon(t)$ is simply

$$\epsilon(t) = \Psi(t) - A\Phi(t). \tag{3.4}$$

Thus the coefficients $A = \{a_i\}$ can be estimated using standard squared-error minimization techniques, such as ordinary least squares (OLS). OLS can be interpreted as the decomposition of $\Psi(t)$ onto the subspace spanned by $\Phi_i(t)$.

Notice that equation (3.3) considers the music features and the emotions as non-causal time series because information about the past (previous times) and about the future (all succeeding times) is used. Equation (3.3) simply models $\Psi(t)$ as a linear combination of a set of feature vectors $\Phi(t)$ where time is treated as vector dimensions. Mathematically, $\Psi(t)$ is projected onto the subspace that $\Phi(t)$ spans, which is usually not orthogonal. This means that the music features used might be linearly dependent. In other words, if one of the features can be expressed as a linear combination of the others, then it is redundant in the feature set because it is correlated (colinear) with the other features.

More importantly, information about the rate of change of musical features is not exploited. The temporal correlation between successive values of features also plays an important role in listeners' emotional experience. The model in equation (3.3) supposes that listeners' emotional responses over time depend on *loudness* values over time, but not on *loudness* "variations". A straightforward way to consider variations in time series is to create a new sequence of values with the first order differences as follows

$$\Delta\Psi(t) = A\Delta\Phi(t) + ... + \epsilon \tag{3.5}$$

where Δ is the first order difference operator $\Delta\Psi(t) = \Psi(t) - \Psi(t-1)$. Difference time series answer questions like "how much does Ψ change when Φ changes"? [36].

Schubert [36] proposed to use music features (loudness, tempo, melodic contour, texture, and spectral centroid) as predictors in linear regression models of valence and arousal. This study found that changes in loudness and tempo

were associated positively with changes in arousal, and melodic contour varied positively with valence. When Schubert [38] discussed modeling emotion as a continuous, statistical function of musical parameters, he argued that the statistical modeling of memory is a significant step forward in understanding aesthetic responses to music. In simple terms, the current system output depends on its previous values. Another interpretation is that the system exhibits "inertia", i.e., no sudden changes occur. Naturally, the input variables (music features) are also likely to exhibit autocorrelation.

Finally, Schubert [37] studied the causal connections between resting points and emotional responses using interrupted time series analysis. This study is related to a hypothesis proposed by Leonard Meyer [25] that "arousal of affect" results from musical expectations being temporarily suspended. Meyer suggests that there is a relationship between musical expectations, tension, and arousal. Schubert concluded that resting points are associated with increased *valence*.

The approach proposed by Schubert implicitly assumes that the relationship between the temporal evolution of music features and the emotional trajectories is linear and mutually independent, discarding interactions between music features. The interactions between musical variables are a prominent factor in music perception and call for joint estimation of coupled music features and modeling of said interactions. Finally, Schubert's approach does not generalize, applying to each piece analyzed.

Linear Dynamical System A linear system models a process where the output can be described as a linear combination of the inputs as in equation (2.2). When the input is a stationary signal corrupted by noise, a *Wiener filter* can be used to filter out the noise that has corrupted the signal. The Wiener filter uses the autocorrelation of input signal and crosscorrelation between input and output to estimate the filter, which can be later used to predict future values of the input.

Linear dynamical systems also model the behavior of the input variable $\Phi(t)$, usually from its past values. The *Kalman Filter* gives the solution to generic *linear state space models* of the form

$$\Phi(t) = A\Phi(t-1) + q(t) \tag{3.6}$$

$$\Psi(t) = H\Phi(t) + r(t) \tag{3.7}$$

where vector $\Phi(t)$ is the state and $\Psi(t)$ is the measurement. In other words, the Kalman filter extends the Wiener filter to nonstationary processes, where the adaptive coefficients of the filter are iteratively (recursively) estimated.

Schmidt and Kim [32–34] have worked on the prediction of time-varying arousal-valence pairs as *probability distributions* using *multiple linear regression*, *conditional random fields*, and *Kalman filtering*. Each music track is described by a time-varying probability distribution from a corpus of annotations they have collected with an online collaborative game [16] from several users. Their first effort [33] to predict the emotion distribution over time simply uses multiple linear regression (MLR) to regress multiple feature windows to these annotations

collected at different times without exploiting the time order or the temporal correlation of the features or the emotions.

Then, Schmidt and Kim [32] modeled the temporal evolution of the music features and the emotions as a linear dynamical system (LDS) such as equation (3.6). The model considers the labels $\Psi(t)$ as noisy observations of the observed music features $\Phi(t)$ and uses a Kalman filter approach to fit the parameters. They compare the results against their previous MLR approach, which considers that each pair feature Φ_i annotation Ψ_i is statistically independent and therefore neglects the time-varying nature of music and emotions. Interestingly, they conclude that a single Kalman filter models well the temporal dependence in music emotion prediction for each music track. However, a mixture of Kalman filters must be employed to represent the dynamics of a music collection.

Later, Schmidt and Kim [34] propose to apply conditional random fields (CRF) to investigate how the relationship between music features and emotions evolve in time. They state that CRF models both the relationships between acoustic data (the music features) and emotion space parameters and also how those relationships evolve over time. CRF is a fully connected graphical model of the transition probabilities from each class to all others, thus representing the link between music features and the annotated labels as a set of transition probabilities, similarly to hidden Markov models (HMM). An interesting finding of this work is that the best performing feature for CRF prediction was MFCC rather than spectral contrast as reported earlier [32]. Schmidt and Kim conclude by speculating that this might be an indication that MFCC provides more information than spectral contrast when modeling the temporal evolution of emotion.

Dynamic Texture Mixture A dynamic texture (DT) is a generative model that takes into account both the instantaneous acoustics and the temporal dynamics of audio sequences [8]. The texture is assumed to be a stationary second-order process with arbitrary covariance driven by white Gaussian noise (i.e., a first-order ARMA model). The model consists of two random variables, an *observed variable* $\Psi(t)$ that encodes the musical emotions, and a *hidden state variable* $\Phi(t)$ that encodes the dynamics (temporal evolution) of the music features. The two variables are modeled as a *linear dynamical system.*

$$\Phi(t) = A\Phi(t-1) + v(t) \tag{3.8}$$

$$\Psi(t) = C\Phi(t) + w(t) \tag{3.9}$$

While the DT in equation (3.8) models a single observed sequence, a *mixture of dynamic textures* (DTM) models a collection of sequences such as different musical features. DTM has been applied in automatic segmentation [2] and annotation [8] of music, as well as MER [44].

Vaizman *et. al* [44] propose to use dynamic texture mixtures (DTM) to "investigate how informative the dynamics of the audio is for emotional content". They created a data set of 76 recordings of piano and vocal performances where

"the performer was instructed to improvise a short musical segment that will *convey to listeners in a clear manner a single emotion*, one from the set of $\{happy, sad, angry, fearful\}$ [44]." These instructions were then used as ground truth labels. Vaizman *et. al* claim that they "obtained a relatively wide variety of acoustic manifestations for each emotional category, which presumably capture the various strategies and aspects of how these specific emotions can be conveyed in Western music." Finally, they model the dynamics of acoustic properties of the music applying DTM to a temporal sequence of MFCCs extracted from their recordings. A different DTM model must be trained for each class (emotional label) using an iterative expectation maximization (EM) algorithm. After training, we can calculate the likelihood that a new music track was "generated" by a given DTM (i.e., the track belongs to that class). Notice that the model in equation (3.8) is equivalent to a first-order state space model.

Auto Regressive Model Korhonen *et al.* [18] assume that, since music changes over time, musical emotions can also change dynamically. Therefore, they propose to measure emotion as a function of time over the course of a piece and subsequently model the time-varying emotional trajectory as a function of music features. More specifically, their model assumes that musical emotions depend on present and past feature values, including information about the rate of change or dynamics of the features. Mathematically, the model has the general form

$$\Psi_i(t, A) = f\left[\Phi_i(t), \Phi_i(t-1), ..., \epsilon_i(t), \epsilon_i(t-1)\right] \tag{3.10}$$

where $\Psi_i(t, A)$ represents the emotions as a function of time t, A are the parameters of the function f that maps the music features $\Phi_i(t)$ and its past values $\Phi_i(t-1), ...$ with approximation error $\epsilon(t)$. Notice that the model does not include dependence on past values of $\Psi_i(t, A)$.

In this work, Korhonen *et al.* [18] adopt linear models, assuming that f can be estimated as a linear combination of current and past music features Φ given an estimation error ϵ to be minimized via least-squares and validated by K-fold cross-validation and statistical properties of the residual error ϵ [18]. The models they consider are the auto-regressive with exogenous inputs (ARX) shown in equation (3.11) and a state-space representation shown in equations (3.12) and (3.13) following.

$$\begin{aligned} \Psi(t) + A_1(\theta)\Psi(t-1) + ... + A_m(\theta)\Psi(t-m) = \\ B_0(\theta)\Phi(t) + ... + B_n(\theta)\Phi(t-n) + e(t) \end{aligned} \tag{3.11}$$

where $\Phi(t)$ is the N-dimensional music feature vector (N is the number of features), $\Psi(t)$ is an M-dimensional musical emotion vector (M is the dimension of the emotion representation), A_k is a matrix of coefficients (zeros) and B_k is the matrix of coefficients (poles).

$$\Phi(t+1) = A(\theta)\Phi(t) + B(\theta)u(t) + K(\theta)\epsilon(t) \tag{3.12}$$

$$\Psi(t) = C(\theta)\Phi(t) + D(\theta)u(t) + \epsilon(t) \tag{3.13}$$

where $\Phi(t)$ is the N-dimensional music feature vector (N is the number of features), $A(\theta)$ is a matrix representing the dynamics of the state vector, $B(\theta)$ is a matrix describing how the inputs (music features) affect the state variables Φ, $C(\theta)$ is a matrix describing how the state variables Φ affect the outputs (emotion), $D(\theta)$ is a matrix describing how the current inputs (music features) affect the current outputs, and $K(\theta)$ is a matrix that models the noise in the state vector Φ. They used a dataset of 6 pieces "to limit the scope," while the total duration was 20 min. They report that the best model structure was ARX using 16 music features and 38 parameters, whose performance was 21.9% for valence and 78.4% for arousal. An interesting conclusion is that previous valence appraisals can be used to estimate arousal, but not the other way around.

Artificial Neural Networks Coutinho and Cangelosi [5–7] propose to use recurrent neural networks to model continuous measurements of emotional response to music. Their approach assumes "that the spatio-temporal patterns of sound convey information about the nature of human affective experience with music" [6]. The temporal dimension accounts for the dynamics of music features and emotional trajectories and the spatial component accounts for the parallel contribution of various musical and psycho-acoustic factors to model continuous measurements of musical emotions.

Artificial neural networks (ANN) are nonlinear adaptive systems consisting of interconnected groups of "artificial neurons" that model complex relationships between inputs and outputs. ANNs can be viewed as nonlinear connectionist approaches to machine learning, implementing both supervised and unsupervised learning. Generally, each "artificial neuron" implements a nonlinear mathematical function $\Psi = f(\Phi)$, such that the output of each neuron is represented as a function of the weighted sum of the inputs as follows

$$\Psi_i = f\left[\sum_j^N w_{ij} g(\Phi_j)\right] \tag{3.14}$$

where Ψ_i is the i^{th} output,Φ_j is the j^{th} input, f is the map between input and output, and g is called *activation function*, usually nonlinear.

There are feed-forward and recurrent networks. Feed-forward networks only use information from the inputs to "learn" the implicit relationship between input and output in the form of connection weights, which act as long-term memory because once the feed-forward network has been trained, the map remains fixed. Recurrent networks use information from past outputs and from the present inputs in a feedback loop. Therefore, recurrent networks can process patterns that vary across time and space, where the feedback connections act as short-term memory (or memory of the immediate past)[3, 6].

Coutinho and Cangelosi [5–7] sustain that the structure of emotion elicited by music is largely dependent on dynamic temporal patterns in low-level music structural parameters. Therefore, they propose to use the Elman neural network (ENN), an extension of feed-forward networks (such as the multi-layer perceptron) that include "context" units to remember past activity by storing and using past computations of the network to influence the present processing. Mathematically,

$$\Phi(t) = f_i\left[\Phi(t-1), u(t)\right] = f\left[\sum_j w_{i,j}\Phi_j(t-1) + \sum_j w_{i,j}u_j(t)\right] \quad (3.15)$$

$$\Psi(t) = h_i\left[\Phi(t)\right] = h\left[\sum_j w_{i,j}\Phi_j(t)\right] \quad (3.16)$$

where equation (3.15) is the *next state function* and equation (3.16) is the *output function*. In these equations, Φ is the musical features, Ψ is the emotion pair $\{\upsilon, \alpha\}$, w are the connection weights (the network long-term memory), and u are the internal states of the network that encode the temporal properties of the sequential input at different levels. The recursive nature of the representation endows the network with the capability of detecting temporal relationships of sequences of features and combinations of features at different time lags [6].

This study used the dataset from Korhonen *et al.* [18]. They concluded that the spatio-temporal relationships learned fro the training set were successfully applied to a new set if stimuli and interpret this as long-term memory, as opposed to the dynamics of the system (associated with short-term memory). The result of canonical correlation analysis revealed that *loudness* is positively correlated with *arousal* and negatively with *valence*, *spectral centroid* is positively correlated with both *arousal* and *valence*, *spectral flux* correlated positively with *arousal*, *sharpness* correlated positively with both *arousal* and *valence*, *tempo* is correlated with high *arousal* and positive *valence*, and finally *texture* is positively correlated with *arousal*. Later, Vempala and Russo [45] compared the performance of a feed-forward network and an Elman network for predicting AV ratings of listeners recorded over time for musical excerpts. They found similar correlations between music features and $\{\upsilon, \alpha\}$ values.

3.4 Overview

This section presents a brief overview of the techniques discussed previously. Table 1 summarizes summarizes features of the models for each approach, providing comments on aspects such as limitations and applicability.

4 Discussion

Most approaches that treat emotional responses to music as a time-varying function of the temporal variation of music features implicitly assume that time

Table 1. Overview of the proposals to model musical emotions from time-varying music features. The table briefly summarizes model features with general comments for each of the approaches reviewed in section 3.3

Approach	Model Features	Comments
Ordinary Least Squares	• Linear • Stationary • Noncausal • Independent estimation • Memoryless	• Does not model temporal system dynamics • Does not model interactions between music features • Models arousal and valence separately • Models each piece separately • Least-squares error minimization
Linear Dynamical System	• Linear • Stationary (Wiener, CRF) • Nonstationary (Kalman) • Causal • Independent estimation (Wiener, Kalman) • Joint estimation (CRF) • Memoryless	• Models temporal system dynamics • Does not model interactions between music features (Wiener, Kalman) • Models each piece separately • Least-squares error minimization • Underlying filtering model is hardly musical
Dynamic Texture Mixture	• Linear • Stationary • Causal • Independent estimation • memoryless	• Models temporal system dynamics • Does not model interactions between music features • Borrowed from video • Models each piece separately • Expectation maximization parameter fit
Auto Regressive Model	• Linear • Stationary • Causal • Independent estimation • Memoryless	• Models temporal system dynamics • Does not model interactions between music features • Borrowed from statistics • Models each piece separately • Least-squares error minimization
Artificial Neural Network	• Nonlinear • Nonstationary • Causal • Joint estimation • Memory	• Models temporal system dynamics • Models interactions between music features • Many parameters • Difficult interpretation

presents certain deterministic properties. In the models discussed above, time is modeled as clock time. However, musical time can be very subjective as music is experienced by the listener. Naturally, listeners' emotional reactions to music are closely related to the subjective experience of time rather than objective clock time. An interesting analogy is perception of frequencies and the Mel scale [43]. Human auditory perception of frequencies is closer to logarithmic rather than linear, thus linear frequency representations such as the Fourier transform present a distorted picture of the information that is used to interpret the sounds that reach the ear. Therefore, in what follows, this article discusses modeling time in MER as subjective musical time rather than objective clock time.

4.1 Time

Conceptually, time can be seen from an objective or subjective point of view. Clocks are evidence of the objective interpretation of time as independent of anyone to experience it. Subjectively, the notion of time comes from the experience of change, sensory or otherwise [29]. Pressing [29] states that "Time is not a stimulus but a construction, an inference." Scientifically, the concept of time can be incorporated into measurements of physical quantities. In this case, time is a measure of change that involves expenditure of energy and therefore increase in entropy. Thus physical time is directly linked to the tendency of macroscopic physical systems to disorder. As a consequence, physical time involves irreversibility on macroscopic scales.

However, musical time differs from scientific time in many respects. Possible procedures to establish the nature of musical time are mathematical formalism and cognitive psychology. Mathematical formalism usually addresses objective clock time, which may be used to model the temporal processes used by composers. Cognitive psychology is concerned with subjective time, studying the mental representation of time.

Newton constructed a deterministic set of mathematical relations that allowed prediction of the future behavior of moving objects and allowed deduction of the past behavior of the moving objects. All that one needed in order to do this was data in the present regarding these moving objects. Isaac Newton believed in absolute space and absolute time. According to the Newtonian view, time is a dimension in which events and objects "move through" or an entity that "flows". Gottfried Leibniz and Immanuel Kant, among others, believed that time and space "do not exist in and of themselves, but ... are the product of the way we represent things", because we can know objects only as they appear to us.

Scientific properties of time Usually, objective time presents some properties as follows [29]

1. *Time provides an ordering for events.* In classical physics and ordinary experience, this ordering is unique for any given set of events and chosen observer.

2. *This ordering has a unique direction.* This unique direction gives rise to the irreversibility of some macroscopic phenomena and is related to the rise in entropy (or disorder) of isolated systems.
3. *Time separates events into three distinct categories: past, present, future.*
4. *Time is measurable.* The existence of clocks that agree to high accuracy (in non-relativistic surroundings) provides the utility of this notion. Clock time is virtually synonymous with scientific time. Time's measurability means that in mathematical terms it acts as a metric space, i.e. a space with a function that defines distance.
5. *Time is continuous (but also discrete).* In classical physics, time is continuous. Quantum mechanics provides a discrete interpretation of time based on the principle of uncertainty.

Musical Properties of Time The properties of scientific time have parallels in music [29]. For example, the musical events have a unique time ordering and the unique *direction* of time is usually "accepted." Also, past, present, and future remain useful concepts, and all musical events are subject to clock measurability. Finally, the continuity or arbitrary divisibility of time applies to sound perception. Most of these properties are associated with objective clock time, such as measured by a metronome or marked on scores. However, when it comes to listening to music, musical time has a subjective, experienced, psychological component. The composer Dennis Smalley [40, 41] wrote that "spectrum is perceived through time and time is perceived as spectral motion", suggesting that sound perception is inherently linked to the auditory perception of change.

Some properties of objective time listed above are modified in musical time. The most affected are 1, 2, and 4. Musically, time is inferred from ordered events. Thus time perception can only be approximately modeled as clock time because we ignore timing differences (and even tempo differences) to a substantial degree. The directionality of time is first of all a property of short-term memory. As for long-term memory, we have a memory of duration, but our memory of time order is rather imprecise once things are in the past. Redundancy is in an interesting way related to the temporal order of musical events and directionality of musical time. Recycling a theme is not just a way of improving long-term memory storage, it is also a musical way of making the time order less important.

The dichotomy between clock (objective) time and experienced (subjective) time has been the subject of considerable debate in music [29]. Snyder [42] views time as linked to the rate of change of incoming information. In this discussion, Snyder wrote that information refers to novelty and the removal of uncertainty. Habituation occurs at many levels of consciousness, cognitive as well as perceptual, and on many different time scales, from seconds to years. Thus we may not notice or remember experiences that keep repeating. However, the limitation of the capacity of memory is a limitation on how much novelty (i.e., *information*) it can handle. To be coherent and memorable, a message must have a certain amount of *non-informative repetition* or *redundancy*, which produces a certain amount of invariance or regularity. The redundancy in messages acts as a kind

of implicit memory rehearsal, allowing us to have certain expectations about the messages we perceive and making them predictable to some extent.

In relation to music, we can find redundancy at different levels of music experience. Repetition of similar waveforms create pitch perception. The concepts of rhythm, tempo, and meter rely on repetition. The constraints of tuning systems and scales limit the number of elements used in a melody, creating redundancy in melodic patterns. At the formal level, redundancy includes symmetries and repetition of entire sections. Snyder suggests that this repetition, in addition to being a memory retrieval cue, is a *metaphor* for the process of remembering itself. When a pattern that appeared earlier in a piece of music reappears, it is like a recollection - an image of the past reappearing in the present, and its familiarity gives stability. Therefore, Snyder proposes that these associative repetitions are a factor in establishing closure, and points that introduce new and unfamiliar material (higher information content), such as transitions, are less stable and have a higher tension value. Snyder concludes that information can be related to tension in music. Musical tension, in turn, is associated with emotional experience. As stated before, the patterns of repetition and expectations in music are directly related to listeners' emotional reactions. One could argue that information measures over time are more suitable to bear a causal relation with *arousal/valence* ratings than music features. But what is the link between the flow of information in music and the perception of time?

Time is often thought of as existing independently of human experience. This objective notion of time is closely related to scientific concept of irreversibility of certain phenomena. Another possible interpretation is that time is an abstract construction of the human mind based on certain aspects of memory. The subjective notion of time is *constructed from* our perceptions of objects and events, and its qualities at a given moment depend on the relationships between these perceptions. In this sense, what we perceive in a given amount of time to some extent determines our sense of the length of that time. In other words, *subjective time* perception is a measure of the flow of information.

The concepts of information and redundancy are intrinsically related to musical form especially because they have a profound effect on our perception and memory of *lengths of time*. Our judgment of the length of a time period longer than the limits of short-term memory depends on the nature of the events that "fill" it. At first, it might seem reasonable to assume that how long a length of time appears to take depends on how many events happen within it, but in reality it seems to depend also on *how much information* we process from those events. Thus a time period filled with novel and unexpected events will be remembered as longer than an identical (in clock time) period filled with redundant or expected events. This implies that our expectations affect our sense of duration. Novel events take up more memory space and are usually remembered as having taken longer. On the other hand, ordinary events,which fit comfortably within our predefined schemas and require little attention and processing, are described as taking up little memory space and in retrospect seem to have taken less time to happen.

Note, however, that the above are descriptions of duration as *remembered*, not as *experienced*. Indeed, duration as experienced tend to be the opposite of duration remembered. "Boring" time periods with little information are experienced as being long, but *remembered as shorter*. Conversely, time periods filled with unusual, informative sequences of events, can seem to flow very rapidly while occurring, but are *remembered as longer*. Thus a musical passage filled with repetitive events can seem, in retrospect, shorter than one filled with unpredictable events. In other words, proportional relations of clock time do not necessarily establish similar relations of proportional *experienced time* or *remembered time* lengths. However, this effect seems to diminish with repeated listening. In addition, regular pulse and metrical frameworks seem to make it easier to get a more accurate sense of larger durational proportions.

Musical time is designed by composer and articulated by performer, shaping the perceptual processes of the listener. Systematic repetition of patterns can dull time perception, stretch or even eliminate the parallels between objective and subjective time. Continuity can be undermined by many traditional musical procedures, such as *staccato*. The hierarchical nature of time is intrinsically related to the three levels of time perception, such that "horizontal" aspects of time focus on on succession of events whereas "vertical" aspects focus on coordination between parts, synchrony, overlay, among others.

5 Conclusions

Research on automatic recognition of emotion in music, still in its infancy, has focused on comparing "emotional labels" automatically calculated from different representations of music with those of human annotators. MER systems commonly use supervised learning techniques to map non time-varying music feature vectors into regions of the emotion space. The music features are typically extracted from short audio clips and the system associates one emotion to each piece. The performance of MER systems using machine learning has been stagnant. Studies in music psychology suggest that time is essential in emotional expression. In this article, we argue that MER has neglected the temporal nature of music. We advocate the incorporation of time in both the representation of musical features and the model of emotions. This article reviews recent proposals in the literature to model musical emotions from time-varying music features. Finally, we discussed the representation of musical time as clock time, rather than subjective time.

The drawbacks of applying supervised learning to non time-varying representations of music and emotions are widely recognized by MER researchers. However, there is no standard way of representing temporal information in MER. This article urges MER researchers to model musical emotions from time-varying music features. The main point we make is that the temporal dynamics of music features are better predictors of musical emotions than feature values. However, we argue that currently, the models that take temporal dynamics into consideration are not appropriate to deal with music because they were originally

developed for other purposes. Currently, we have the means to model the relevant features over scientific (clock) time. However, musical time is not in the equation.

Future perspectives include the development of computational models that exploit the temporal dynamics of music features as predictors of musical emotions. Only by including temporal information in automatic recognition of emotions can we advance MER systems to cope with the complexity of human emotions in one of its canonical means of expression, music.

References

1. Aucouturier, J.J., Pachet, F.: Improving Timbre Similarity: How High is the Sky? Journal of Negative Results in Speech and Audio Sciences, 1(1) (2004).
2. Luke Barrington and Antoni B. Chan and Gert Lanckriet: Dynamic texture models of music. In: Proc. ICASSP, (2009).
3. Caetano, M., Wiering, F.: The Role of Time in Music Emotion Recognition. In: Proceedings of the International Symposium on Computer Music Modeling and Retrieval (2012).
4. Celma, O., Serra, X.: FOAFing the Music: Bridging the Semantic Gap in Music Recommendation. Journal of Web Semantics, 6(4) (2008).
5. Coutinho, Eduardo; Cangelosi, Angelo. The use of spatio-temporal connectionist models in psychological studies of musical emotions. Music Perception: An Interdisciplinary Journal, 27(1), pp. 1–15 (2009).
6. Coutinho, E., Cangelosi, A. A Neural Network Model for the Prediction of Musical Emotions. In S. Nefti-Meziani, J.G. Grey (Ed.), Advances in Cognitive Systems, pp. 331–368, London: IET Publisher (ISBN: 978-1849190756) (2010).
7. Coutinho E., Cangelosi A.: Musical emotions: predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. Emotion, 11(4), pp. 921–37 (2011).
8. Coviello, E. and Chan, A.B. and Lanckriet, G.: Time Series Models for Semantic Music Annotation. IEEE Transactions on Audio, Speech, and Language Processing. 19 (5), pp. 1343–1359 (2011).
9. Gabrielsson, A., Lindström, E.: The Role of Structure in the Musical Expression of Emotions. In: Handbook of Music and Emotion: Theory, Research, Applications. Eds. Patrik N. Juslin and John Sloboda, pp. 367–400 (2011)
10. Hevner, K.: Experimental Studies of the Elements of Expression in Music. The Am. Journ. Psychology. 48 (2), pp. 246–268 (1936)
11. Hu, X., Downie, J.S., Laurier, C., Bay, M., and Ehmann, A.F.: The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In: Proc. ISMIR (2008)
12. Huq, A., Bello, J.P., and Rowe, R.: Automated Music Emotion Recognition: A Systematic Evaluation. Journal of New Music Research. 39(4), pp. 227–244 (2010)
13. Huron, D.: Sweet Anticipation: Music and the Psychology of Expectation. MIT Press, (2006)
14. Juslin, P. N., Västfjäll, D.: Emotional Responses to Music: The Need to Consider Underlying Mechanisms. Behavioral and Brain Sciences, 31(5), pp. 559–621 (2008).
15. Juslin, P., Timmers, R.: Expression and Communication of Emotion in Music Performance. In: Handbook of Music and Emotion: Theory, Research, Applications Eds. Patrik N. Juslin and John Sloboda, pp. 453–489 (2011)

16. Y. Kim, E. Schmidt, and L. Emelle. MoodSwings: A. Collaborative Game for Music Mood Label Collection. In: Proceedings of the 9th International Conference on Music Information Retrieval ISMIR, (2008)
17. Kim, Y.E., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., Speck, J., Turnbull, D.: Music Emotion Recognition: A State of the Art Review. In: Proc. ISMIR (2010)
18. Korhonen, M.D., Clausi, D.A., Jernigan, M.E.: Modeling Emotional Content of Music Using System Identification. IEEE Transanctions on Systems, Man, and Cybernetics, Part B, Cybernetics. 36(3), pp. 588–599 (2005)
19. Krumhansl, C. L.: An Exploratory Study of Musical Emotions and Psychophysiology. Canadian Journal of Experimental Psychology. 51, pp. 336–352 (1997)
20. Krumhansl, C. L.: Music: A Link Between Cognition and Emotion. Current Directions in Psychological Science. 11, pp. 45–50 (2002)
21. Lu, L., Liu, D., Zhang, H-J.: Automatic Mood Detection and Tracking of Music Audio Signals. IEEE Trans. Audio, Speech, Lang. Proc. 14, 1 (2006)
22. McAlpin, C.: Is Music the Language of Emotions? The Musical Quarterly, 11(3), pp. 427–443 (1925).
23. MacDorman, K. F., Ough S., Ho C.C.: Automatic Emotion Prediction of Song Excerpts: Index Construction, Algorithm Design, and Empirical Comparison. Journal of New Music Research. 36, pp. 283–301 (2007)
24. Meyer, L. B.: Music, the Arts, and Ideas. University of Chicago Press, Chicago (1967)
25. Meyer, L. B.: Emotion and Meaning in Music. University of Chicago Press, Chicago (1956)
26. Mion, L., Poli, G.: Score-Independent Audio Features for Description of Music Expression. IEEE Transanctions on Audio, Speech, and Language Processing, 16(2), pp. 458–466 (2008).
27. Müller, M., Ellis, D.P.W., Klapuri, A., Richard, G.: Signal Processing for Music Analysis. IEEE Journal of Selected Topics in Sig. Proc. 5(6), pp. 1088–1110 (2011)
28. Nagel, F., Kopiez, R., Grewe, O., Altenmüller, E.: EMuJoy. Software for the Continuous Measurement of Emotions in Music. Behavior Research Methods, 39 (2), pp. 283–290 (2007)
29. Pressing, J.: Relations Between Musical and Scientific Properties of Time. Contemporary Music Review, 7(2), pp. 105–122 (1993).
30. Russell, J.A.: A Circumplex Model of Affect. Journ. Personality and Social Psychology. 39, pp. 1161–1178 (1980)
31. Scherer, K.R. Which Emotions Can be Induced by Music? What are the Underlying Mechanisms? and How can We Measure Them?. Journal of New Music Research, 33(3), pp. 239-251, (2005).
32. Schmidt, E.M., Kim, Y.E.: Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering. In: Proc. Ninth International Conference on Machine Learning and Applications (ICMLA), pp.655-660, (2010)
33. Schmidt, E.M., Kim, Y.E.: Prediction of Time-Varying Musical Mood Distributions from Audio. In: Proc. ISMIR (2010)
34. Schmidt, E.M., Kim, Y.E.: Modeling Musical Emotion Dynamics with Conditional Random Fields. In: Proc. ISMIR (2011)
35. Schubert, E.: Measuring Emotion Continuously: Validity and Reliability of the Two-Dimensional Emotion Space. Australian Journal of Psychology, 51(3), pp.154–165 (1999).
36. Schubert, E.: Modeling Perceived Emotion with Continuous Musical Features. Music Perception, 21(4), pp. 561–585 (2004)

37. Schubert, E.: Introduction to Interrupted Time Series Analysis of Emotion in Music: The Case of Arousal, Valence and Points of Rest. In: Proc. International Conference on Music Perception and Cognition (2004)
38. Schubert, E.: Analysis of Emotional Dimensions in Music Using Time Series Techniques. Journ. Music Research, 31, pp. 65–80 (2006).
39. Schubert, E.: Continuous Self-Report Methods. In: Handbook of Music and Emotion: Theory, Research, Applications Eds. Patrik N. Juslin and John Sloboda, pp. 223–254 (2011)
40. Smalley, D.: Spectro-morphology and Structuring Processes. In: Emmerson, S. (ed.) The Language of Electroacoustic Music. London: Macmillan. pp 61–93 (1986).
41. Smalley, D.: Spectromorphology: Explaining Sound-Shapes. Organised Sound 2 (2), pp. 107–126 (1997).
42. B. Snyder: *Music and Memory: An Introduction*, MIT Press (2001).
43. Stevens, S. S., Volkman, J., Newman, E. B.: A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America 8 (3), pp. 185–190 (1937).
44. Vaizman, Y., Granot, R.Y., Lanckriet, G.: Modeling Dynamic Patterns for Emotional Content in Music. In: Proc. ISMIR (2011)
45. Naresh N. Vempala and Frank A. Russo: Predicting Emotion from Music Audio Features Using Neural Networks. In: Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR), (2012).
46. Vos P G, van Dijk A, Schomaker L.: Melodic Cues for Metre. Perception, 23(8), pp. 965–976 (1994).
47. Wiggins, G. A.: Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music. IEEE International Symposium on Multimedia, pp. 477-482 (2009)
48. Yang, Y. H., Lin, Y. C., Su, Y. F., Chen, H. H.: A regression approach to music emotion recognition. IEEE Trans. Audio, Speech, Lang. Proc. 16 (2) pp. 448–457 (2008)
49. Yang, Y., Chen, H.: Ranking-Based Emotion Recognition for Music Organization and Retrieval. IEEE Trans. Audio, Speech, Lang. Proc. 19, 4 (2011)